PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 15/16, 17/00, 17/30 — A1

(11) International Publication Number: WO 95/21415

(43) International Publication Date: 10 August 1995 (10.08.95)

(21) International Application Number: PCT/US95/01566

(22) International Filing Date: 7 February 1995 (07.02.95)

(30) Priority Data:

08/192,654	7 February 1994 (07.02.94)	US
08/246,246	19 May 1994 (19.05.94)	US

(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

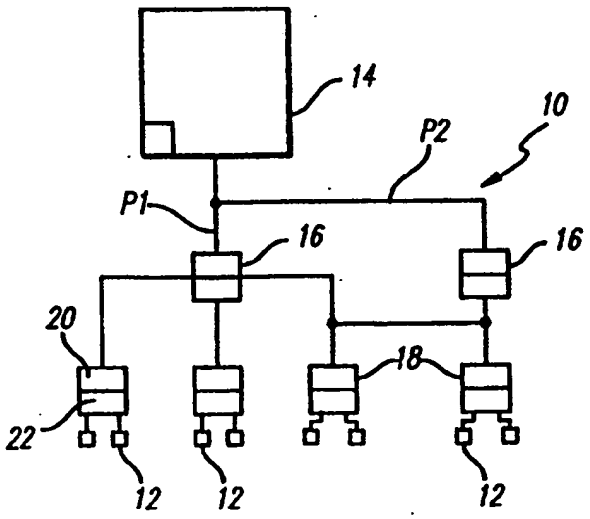
Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(71) Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 22nd floor, 300 Lakeside Drive, Oakland, CA 94612-3550 (US).

(72) Inventors: PAPADIMITRIOU, Christos; Department of Computer Sciences & Engineering, AP & M, Room 4161, University of California at San Diego, La Jolla, CA 92093-0114 (US). RANGAN, P., Venkat; 13011 Callcott Way, San Diego, CA 92130 (US).

(74) Agent: BERLINER, Robert; Robbins, Berliner & Carson, 5th floor, 201 N. Figueroa Street, Los Angeles, CA 90012 (US).

(54) Title: SYSTEM FOR MULTIMEDIA INFORMATION DELIVERY



(57) Abstract

A multimedia information delivery network system (10) is disclosed for delivering multimedia programs to a plurality of users (12) at user-selected times. The network includes a wide area transmitter (14) for transmitting the multimedia programs. Additionally, the network includes a plurality of network servers (16) for receiving the programs and for selectively caching the programs for retransmission to downstream network servers (18) and/or directly to one or more users (12) at the user-selected transmission times. A scheduler (36) receives the user-selected transmission times and, in response thereto, establishes a network server path by which the multimedia program is efficiently delivered to each user (12) at the respective user-selected time.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

SYSTEM FOR MULTIMEDIA INFORMATION DELIVERY

FIELD OF THE INVENTION

The present application is a continuation-in-part of, and claims priority from, co-pending U.S. patent application serial number 08/192,654, filed February 7, 1994, for an invention entitled "SYSTEM FOR EFFICIENT DELIVERY OF MULTIMEDIA INFORMATION", the disclosure of which is incorporated herein by reference.

The present invention relates generally to multimedia system architectures, and more particularly to methods and systems for efficient, on-demand delivery of multimedia information and programs.

BACKGROUND

Historically, both educational information and entertainment programs have been generated by content providers and delivered to users in a variety of formats, including printed publications, data storage media, and electronic broadcast signals. To access printed publications and data storage media, users are required to borrow or purchase the publications or media. This typically requires the users who desire access to travel to the information storage facilities to obtain the desired media, and in any case to store the media at the users' facilities for access at a time convenient for the users.

On the other hand, gaining access to information carried by broadcast signals is comparatively simpler, requiring only that the user tune a receiver in to the broadcast at the scheduled broadcast time. Unfortunately, the information sought by the user is available only at the time of broadcast, which may not necessarily be the time most convenient to the user.

Nevertheless, recognizing the advantages inherent in electronically broadcasting information, many content providers now electronically publish educational information for broadcast which was previously made available only in printed publications. A trend is therefore clearly emerging, colloquially termed the "information superhighway", in which large amounts of educational and entertainment programs and information ("multimedia") will be electronically available to a large number of users. Stated differently, multimedia encompasses both traditional audio-video television-type programs and programs spawned by the convergence of diverse industries and technologies including communications, computer, entertainment, publishing, film, and television, and multimedia programming is expected to be widely available.

Under certain circumstances, a user may be able to record a broadcast program for review at a later, more convenient time than the broadcast time. It will be appreciated, however, that requiring each user to store a particular program on a storage medium, e.g., a VCR,

dedicated only to that user is inefficient. More particularly, each user must obtain his or her own storage medium and record the program separately from the storage media of other users, thereby duplicating resources and effort and consequently greatly increasing the overall cost to the users who wish to access the program.

Not surprisingly, attempts have been made to address the problem noted above by providing multimedia architectures which permit so-called "video-on-demand". Most of these architectures share the common attribute of requiring immense central data storage servers capable of storing vast amounts of data. These architectures necessarily entail unprecedented demands on the networks over which the data from the storage servers can be transmitted to users.

Unfortunately, a single motion picture requires several gigabytes of storage space. Accordingly, it will be appreciated that storing a full library of motion pictures, a full library of digitized literature and imagery, encyclopedic data, and so on, and, in turn, broadcasting such information on demand to various users at various times around the clock, would stretch and exceed the capacity of current storage and network technology.

In light of the above, it is an object of the present invention to provide a multimedia architecture for efficiently delivering video at user-defined times.

Another object of the present invention is to provide a hierarchical multimedia delivery architecture which is easy to use and cost-effective.

SUMMARY OF THE INVENTION

A system for delivering a multimedia program to a plurality of users includes at least one wide area server for transmitting the multimedia program. Also, the system includes a plurality of network servers for receiving the multimedia program from the wide area server and for selectively caching the multimedia program for respective storage time periods. The network servers transmit the multimedia program at the expirations of their respective storage time periods to the users. In one special case, at least some of the network servers are all arranged in tandem.

Typically, the transmission of a program from a server has an associated predetermined transmission factor. Further, each storage time period has an associated predetermined caching factor which increases over time, and each storage time period also defines an associated preselected end time. The transmission factors and caching factors associated with delivering the program from the metropolitan area server to each of the users at the respective preselected times establishes an aggregate delivery factor. A scheduler is provided for receiving the preselected times from the users and

for establishing the storage time periods in response thereto to minimize the aggregate delivery factor.

In accordance with the present invention, the scheduler also determines which network servers cache the program for each user. Additionally, the scheduler causes a network server to cache the multimedia program only at one of the preselected times. Moreover, the scheduler prevents a downstream network server from transmitting the program back to the wide area server.

In another aspect of the present invention, a multimedia delivery network for delivering a multimedia program to a plurality of users has a network architecture that is characterized by at least one wide area multimedia program transmitter. Also, the network includes at least one network switch in tandem with the wide area multimedia program transmitter for relaying the multimedia program to a user. Furthermore, the network includes a network storage device which is associated with the network switch for caching the multimedia program for a storage time period and for causing the switch to retransmit the multimedia program to a user at a preselected time.

In still another aspect of the present invention, a method is disclosed for delivering a multimedia program to a plurality of users at user-selected times. The method of the present invention includes the steps of transmitting the program from a wide area transmitter, and receiving the program at a plurality of network

servers. Then, each network server is caused to retransmit the program to the end users at one or more of the respective preselected times. Preferably, the method further includes causing one or more network servers to cache the program for an associated storage time period. For optimal performance, a downstream network server would tap the program as it flows from an upstream network server to the end user.

In an alternate embodiment, for a multimedia system which includes a plurality of program servers, a method is disclosed for delivering a multimedia program which has a primary transmission time to first and second users at respective user-selected first and second times, wherein the first time is simultaneous with or subsequent to the primary transmission time and the second time is subsequent to the first time. The method includes the steps of, for each server, determining a first transmission factor that is associated with transmitting the program from the server to the first user. Then, for each server, a first storage factor which is associated with storing the program at the server from the primary transmission time to the first time is determined. Then, the server with the lowest combined first transmission factor and first storage factor is designated as the first server for transmitting the program to the first user at the first time.

Next, for each server, a second transmission factor which is associated with transmitting the program from

the server to the second user is determined. Also, for each server, a second storage factor which is associated with storing the program at the server from the first time to the second time is determined. The server having the lowest combined second transmission factor and second storage factor is designated the second server for transmitting the program to the second user at the second time.

In another aspect of the alternate embodiment, a multimedia system is disclosed for delivering a multimedia program having a primary transmission time to first and second users at respective user-selected first and second times. The first time is simultaneous with or subsequent to the primary transmission time and the second time is subsequent to the first time, and the system of the present invention includes a plurality of program servers.

The system further includes means for determining, for each server, a first transmission factor associated with transmitting the program from the server to the first user, and means for determining, for each server, a first storage factor associated with storing the program at the server from the primary transmission time to the first time. Further, the system includes means for designating as a first server the server having the lowest combined first transmission factor and first storage factor for transmitting the program to the first user at the first time.

Additionally, the system includes means for determining, for each server, a second transmission factor associated with transmitting the program from the server to the second user, and means for determining, for each server, a second storage factor associated with storing the program at the server from the first time to the second time. In accordance with the present invention, the system also includes means for designating as a second server the server having the lowest combined second transmission factor and second storage factor for transmitting the program to the second user at the second time.

In still another aspect of the alternate embodiment, a method is disclosed for delivering a multimedia program to a plurality of users at respective user-defined viewing times from a network having a plurality of servers, with each server including a cache for storing the program with at least one of the servers being in the transmission path between another server and a user. The method includes, for the user having the second-earliest associated viewing time, determining a transmission factor associated with transmitting the program from the server to the user for each server that can be used as a cache for the user. Further, the method includes determining the lower of a first test storage factor associated with storing the program at the server from the last time the server was used as a cache until the associated viewing time and a second test storage factor

associated with storing the program at the server from the last time the server was on a transmission path to a previous user until the associated viewing time, for each server that can be used as a cache for the user.

Next, the lower of the first test storage factor and second test storage factor is selected as a storage factor, and the server that has the lowest combined transmission factor and storage factor is selected for transmitting the program to the user at the associated viewing time. The process described above is repeated for each subsequent user, in chronological order of the associated viewing times.

The details of the present invention, both as to its structure and operation, can best be understood in reference to the accompanying drawings, in which like reference numeral refer to like parts, and in which:

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic diagram showing a hierarchical multimedia network architecture;

Figure 2 is a schematic diagram showing a simplified multimedia network architecture using only a single path of tandem servers;

Figure 3 is a graphical representation of the network shown in Figure 2;

Figure 4 is a flow chart showing the operational steps of the network shown in Figure 2;

Figure 5 is a flow chart showing the operational steps of a personal service agent; and

Figure 6 is a flow chart showing an alternate embodiment of the operational steps of the network shown in Figure 2.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring initially to Figure 1, a hierarchical multimedia network is shown, generally designated 10. In accordance with the present invention, the network 10 transmits video-based programs, including multimedia educational and entertainment programs that may contain alpha-numeric displays in addition to video displays and audio signals, to a plurality of users 12. Stated differently, the present invention encompasses the delivery of both traditional audio-video television-type programs, as well as multimedia programs spawned by the convergence of diverse industries and technologies including communications, computer, entertainment, publishing, film, and television. It is to be understood that the network 10 transmits its video-based programs via fiber-optics communication lines and other land lines known in the art, satellite broadcast, personal wireless communications, and/or some combination of the above.

As the skilled artisan will appreciate, the users 12 include personal or local receiving apparatus, e.g., personal television sets or computers. Each user 12 may be linked to the network 10 by land lines or via personal

wireless communications. Also, as more fully disclosed below, each user 12 may include automated means for selecting particular multimedia programs transmitted by the network 10, and for selecting the times at which the user 12 wishes to view the particular programs.

As shown in Figure 1, the network 10 includes at least one wide area server 14. The wide area server 14 may be any appropriate wide area multimedia transmitter, e.g., the equipment associated with transmitting the signals of a network news company.

Also, the network 10 includes at least one and preferably a plurality of primary network servers 16. As shown in Figure 1, each primary network server 16 is in tandem with the wide area server 14. In other words, each primary network server 16 receives the signals transmitted by the wide area server 14. In one presently preferred embodiment, each primary network server 16 is a metropolitan area server, e.g., the equipment associated with transmitting the signals of a metropolitan cable television company.

Further, a plurality of secondary network servers 18 are each arranged in tandem with an associated primary network server 16 for receiving signals therefrom, and as shown in Figure 1 one primary network server 16 can feed signals to one or more secondary network servers 18. In turn, a secondary network server 18 can feed signals to other secondary network servers (not shown). Thus, the primary servers 16 are upstream from their associated

secondary servers 18, and downstream from the wide area server 14. In other words, a downstream network server can tap the program as it flows from an upstream network server to the end user. Each secondary network server 18 can be a multimedia hub, e.g., a neighborhood electronic signal transmission and storage facility.

While Figure 1 shows two levels of network servers 16, 18, it is to be understood that the network 10 can include additional levels of network servers (not shown).

Each network server 16, 18 (and any additional levels of network servers) includes an associated switch 20 and an associated storage device 22. Each switch 20 can advantageously be a network toll switch as is well-known in the art. Thus, each network toll switch 20 essentially functions as a relay by retransmitting, at a preselected time to the server 16, 18 (or user 12) that is immediately downstream of the particular switch 20, the signal received from the server 14, 16, or 18 that is immediately upstream of the server associated with the particular switch 20. The preselected time for retransmission may of course be instantaneously upon receipt of the signal from the upstream server 16, 18.

Moreover, each storage device 22 can be a well-known electronic media storage device, e.g., a magnetic or optical disc with associated support equipment. Accordingly, each storage device 22 can electronically copy, i.e., store, predetermined programs which are received at the associated server 16, 18 for subsequent

retransmission of the programs by the associated switch 20.

Figure 2 shows a simplified multimedia network, generally designated 30, having a wide area server S_1 and a series of network servers S_2 - S_m arranged in tandem for delivering multimedia programs to a plurality of users U_1 - U_n at respective preselected delivery times t_1 - t_n. Each transmission from a server S_{k-1} to a server S_k, $k \in (1,m)$ entails an associated transmission factor $NC_{k-1,k}$ which is ordinarily a constant value.

Also, an associated storage factor SC_k is associated with the storage of a program at a server S_k, and each storage factor SC_k increases linearly in proportion to the time period during which the associated program is stored at the server S_k. Stated differently, each storage factor SC_k is essentially the product of a fixed storage coefficient SC_{kcoef} and a storage time period. Together, the aggregate transmission factors $\sum_{k=1}^{n} NC_{k-1,k}$ and storage factors $\sum_{k=1}^{n} SC_k$ establish an overall aggregate delivery factor.

As intended by the present invention, the overall efficiency of the network 30, as measured by the aggregate delivery factor, is minimized by judiciously selecting which servers S store and retransmit the program to which users U as per the following disclosure. It is to be understood that the factors discussed above may in practice be expressed in units such as time,

bandwidth, memory capacity, transmission power, and even monetary cost.

Figure 3 is a graphical representation of the network 30 shown in Figure 2, wherein the network 30 has been mapped to a two dimensional grid 32 which shows the various ways in which a program can be delivered to the users U_1 - U_n. As shown, the ordinate of the grid 32 represents the servers S_1 - S_m, and the abscissa represents the preselected viewing times t_1 - t_n in increasing temporal sequence.

Accordingly, the skilled artisan will appreciate that the vertical lines within the grid 32 represent and are proportional to the transmission factors NC and the horizontal lines represent and are proportional to the storage factors SC. Thus, by minimizing the total length of the lines extending between the 0, S_1 vertex and the t_n, S_m vertex, the aggregate delivery factor can be minimized.

In the network 30 architecture represented in Figures 2 and 3, the aggregate delivery factor can be minimized by selecting a path with the following constraints: The path must be rooted at the 0, S_1 vertex and extend to the t_n, S_m vertex; the path of the program to the n^{th} user (denoted by the arrows 34 in Figure 3) must extend downwardly and rightwardly from the 0, S_1 vertex, and cannot extend leftwardly (or back to the wide area server) during any portion of the delivery; and the program is transmitted from one server S_{a-1} to its

downstream server S_a (or to a user U) only at one of the preselected viewing times t_1 - t_n. Stated differently, the program is transmitted from an upstream server S_{a-1} to its downstream server S_a only when the program is to be transmitted from the upstream server S_{a-1} to a user U, and at no other times.

Figure 4 is a flow chart showing the operational steps of the network 30 shown in Figures 2 and 3. It can be shown for the simplified tandem network 30 shown which uses only a single path of servers S, the below-described method optimally minimizes the aggregate delivery factor of the network 30. It is to be understood that the analysis disclosed below can be performed by a scheduler 36 (e.g., a type 486 personal computer) which is located at the server S_1 or any of the other servers S. Or, the analysis described below can be performed by a scheduler that is a personal service agent (PSA) 38 associated with a respective user U. The PSA 38 is described more fully below.

Starting at block 40 in Figure 4, the scheduler 36 initializes certain parameters to zero. These parameters include all optimal caching schedule factors designated herein as HOCS, VOCS, HVOCS, and OCS.

It is to be understood that $[S_i][t_j, t_r]$ represents a grid having its leftmost vertex at (t_j, S_i) and its rightmost vertex at (t_r, S_m). As intended by the present invention, S_i is the optimal server at which to cache a program for delivery to the last user U_r (i.e., the user

requesting program delivery at a time subsequent to program delivery to all other users) at time t_r, S_m is the last server in the tandem series of servers, and t_j is the optimal time at which the optimal server S_i caches the program.

In accordance with the present invention, $OCS[S_i][t_j, t_r]$ represents the optimum caching schedule factor for the grid $[S_i][t_j, t_r]$. $HOCS[S_i][t_j, t_r]$ represents the optimum caching schedule factor for the grid $[S_i][t_j, t_r]$ under the constraint that the program is being cached at server S_i during the entire time period t_j to t_r. Such a constraint is necessary to ensure optimization. Also, none of the servers upstream of S_i can be used in this schedule, in order to avoid contradicting the definition of time t_j as being the optimal caching time for S_i.

Moreover, $VOCS[S_i, S_k][t_j, t_r]$ represents the optimum caching schedule factor for the grid $[S_i][t_j, t_r]$ under the constraint that one of the servers S_k, $k \in [1,i]$ is used for servicing the last user U_r. Furthermore, $HVOCS[S_i, S_k][t_j, t_r]$ represents the optimum caching schedule factor for the grid $[S_i][t_j, t_r]$ under the constraints that one of the servers S_k is used for servicing the last user U_r, and that the program is cached at server S_i during the period t_j to t_r.

Blocks 42-54 represent a nested series of iterations that are next performed by the scheduler 36. Specifically, block 42 indicates that each path P_i in a multi-tandem hierarchical network is considered (Figures

2 and 3 represent a tandem network 30 which has only a single path, whereas Figure 1 represents a network 10 which has two paths P_1, P_2). Then, as indicated at block 44, for each path the scheduler 36 receives the delivery times t_1 - t_r from the respective users U_1 - U_n, and constructs the grid 32.

Block 46 indicates that the below-described steps are completed for each user request "r", from r_1 to r_n at respective delivery times t_1 to t_r. Next, block 48 indicates that for each user request r, the below-described steps are completed for each server S_i. For each server S_i, block 50 indicates that the below-described steps are completed for each time t_j, from t_1 to t_r. Thus, for any rectangular grid $[S_i][t_j, t_r]$ there are $(t_j - t_r) * (S_m - S_i)$ sub-grids for which horizontal and vertical constraints must determined.

Also, block 52 indicates that for each time t_j, the below-described steps are completed for each server S_a in S_i to S_m. In other words, there are only $(S_m - S_i)$ possible ways in which the program can reach the client r.

Further, block 54 indicates that for each server S_a, the below-described steps are completed for each time t_b in t_j to t_{r-1}. As intended by the present invention, t_b is the time at which the server S_a caches the program.

Next, the factor (V_{fac}) associated with the vertical constraints of the sub-grid $[S_i][t_j, t_r]$ is determined by the scheduler 36 at block 56 as follows:

$V_{fac} = VOCS[S_i, S_a][t_j, t_b] + HOCS[S_a][t_b, t_{r-1}] + SC_{acoef}*(t_r-t_b) + \Sigma_{s=a}^{m-1}NC_{s,s+1}$.

After determining V_{fac}, the scheduler 36 moves to block 58, wherein the scheduler 36 determines whether the server S_a under consideration is the S_i under consideration. If it is, the factor (H_{fac}) associated with the horizontal constraints of the sub-grid $[S_i][t_j, t_r]$ is determined at block 60 as follows:

$H_{fac} = HVOCS[S_i, S_a][t_j, t_b] + HOCS[S_a][t_b, t_{r-1}] + SC_{acoef}*(t_r-t_b) + \Sigma_{s=a}^{m-1}NC_{s,s+1}$.

Otherwise, because caching at S_i does not entail a storage factor, H_{fac} is determined at block 62 as follows:

$H_{fac} = HVOCS[S_i, S_a][t_j, t_b] + HOCS[S_a][t_b, t_{r-1}] + \Sigma_{s=a}^{m-1}NC_{s,s+1}$.

After determining V_{fac} and H_{fac}, the scheduler 36 iteratively repeats the above-described steps for each t_b associated with each particular S_a, and then for the remainder of the possible servers S_a with associated caching times t_b. More particularly, the scheduler 36 moves back to block 54 to retrieve the next t_b, and then repeats blocks 56-62 as appropriate. Upon using the last t_b, the scheduler 36 moves to block 52 to retrieve the next server S_a, and repeats blocks 54-62 for all times t_b for that server S_a.

Each V_{fac} and H_{fac} so determined is stored, and then at block 64, the optimal schedules VOCS, HVOCS, OCS, and HOCS are determined as follows:

$VOCS[S_i, S_a][t_j, t_r]$ = minimum V_{fac} determined above among the various V_{facs} for all servers S_c in the range S_i to S_a, with associated caching times t_b for each server S_c in the range of user request times from t_j to t_{r-1}.

Also,

$HVOCS[S_i, S_a][t_j, t_r]$ = minimum H_{fac} determined above among the various H_{facs} for all servers S_c in the range S_i to S_a, with associated caching times t_b for each server S_c in the range of user request times from t_j to t_{r-1}.

Further,

$OCS[S_i][t_j, t_r]$ = minimum $VOCS[S_i, S_a][t_j, t_r]$ for all servers S_a in the range S_i to S_m; and

$HOCS[S_i][t_j, t_r]$ = minimum $HVOCS[S_i, S_a][t_j, t_r]$ for all servers S_a in the range S_i to S_m.

After block 64, the scheduler 36 moves to block 66, wherein it is determined whether the last t_j has been processed. If not, the scheduler 36 loops back to block 50 to retrieve the next t_j, and then repeats blocks 52-66. If the last t_j has been processed, the scheduler 36 loops back to block 48 to retrieve the next S_i, and then repeats blocks 50-66. If the last S_i has been processed, the scheduler 36 loops back to block 46 to retrieve the next user request, and then repeats blocks 48-66.

For the single-path tandem network 30 shown in Figures 2 and 3, the overall optimum path $OCS[S_1][t_1, t_n]$ resulting from the above steps defines the program

transmission and caching paths to each user U_n. The output of the above steps is accordingly a series of transmission paths and storage server designations.

For the multi-path hierarchical network 10 shown in Figure 1, after processing all user requests for path P_1, and obtaining and storing a corresponding $OCS[S_1][t_1, t_n]$, the scheduler loops back to block 40, reinitializes all VOCS, HOCS, HVOCS, and OCS to zero, and then steps through blocks 42-66 as described for the next path P_2. The minimum $OCS[S_1][t_1, t_n]$ so obtained is then displayed or otherwise output at block 68.

Additionally, in cross-reference to Figures 2 and 5, each user U_1 - U_n can include a respective PSA 38 (only one PSA shown in Figures 2 and 5). Each PSA 38 is preferably a personal computer which analyzes the content of programs previously selected by the user's client (i.e., by the particular person being served) to select future programs for viewing by the client. Also, the PSA 38 schedules the selected programs for delivery to the client at the client's preferred viewing times.

Thus, when a PSA 38 is provided, the preselected delivery time t_k discussed above is established by either the client of the user U or automatically, by the associated PSA 38. Moreover, in lieu of or in addition to the scheduler 36, the PSA 38 can optimize the transmission and storage factors borne by its associated user by adhering to the operational steps outlined above and shown in Figure 4.

As shown at block 70 in Figure 5, the PSA 38 initially analyzes the content of the programs selected by the client of the associated user U to generate inferences regarding the client's preferences. These inferences can depend on the time of day, i.e., the PSA 38 may note that the client prefers news formats in the early morning and educational programming in the evening.

Also, the PSA 38 can generate inferences of client preferences based direct client input regarding preferences. Based upon the inferences generated at block 70, the PSA 38 moves to block 72 to query program providers to determine the availability of programs relevant to the client's preferences. At block 74, the PSA 38 discards programs which are completely uninteresting to the client.

Then, at block 76, the PSA 38 examines the content of the remaining programs to determine which programs are to be delivered to the client by e.g., the network 10 or the network 30. At block 78, the PSA 38 determines when delivery is to occur (i.e., the PSA 38 determines t_r). Next, at block 80, the PSA 38 sends t_r to the scheduler 36 for operation as disclosed in Figure 4. Alternatively, the PSA 38 itself, in conjunction with other PSAs, can establish the delivery schedule in accordance with the method shown above.

Now referring to Figure 6, an alternate embodiment of the operational method of the present invention is shown for use in conjunction with, e.g., the network 10

shown in Figure 1. It is to be understood that the method shown in Figure 6 is used to deliver a program having a primary transmission time to "n" users U_j, $j \in \{1,n\}$ at associated delivery times t_j using a network having "m" servers S_j, $j \in \{1,m\}$, wherein t_1 is the earliest requested time and corresponds to user U_1, and wherein at least one of the servers S_j is in the transmission path between another server S_k, $k \in \{1,j-1\}$ and a user U_j. It is to be further understood that t_1 is simultaneous with or subsequent to the primary transmission time of the program, with each subsequent request time t_i, $i \in \{2,n\}$ being chronologically later than t_{i-1}.

Commencing at step 82, the process begins, and moves to block 84. At block 84, the method considers all users U_j in chronological order of their associated user request times t_j. Thus, for the first iteration of the operational method shown in Figure 6, the method considers the first user U_1, i.e., the user who has requested a program be delivered at some associated time t_1.

Next, at block 86, for each user U_j, the operational method considers each server S_j that can cache a program for the user U_j being considered. Thus, at block 86 during the first iteration of the method shown in Figure 6, the method considers the first server S_1. At block 88, the method determines the transmission factor associated

with transmitting the program from the server S_j to the user U_j in accordance with the principles disclosed above.

Next, at block 90, the method determines the storage factor associated with caching the program at the server S_j since the last time t_b the server S_j was used as a cache or was on a transmission path to a previous user U_{j-1} until the delivery time t_j. In so determining the storage factor, the method uses the below equation:

$$\text{storage factor} = SC_{jcoef} * (t_j - t_b)$$

wherein SC_{jcoef} = fixed storage coefficient for the server S_j.

Stated differently, at block 90 the method determines the lower of a first test storage factor that is associated with storing the program at the server S_j from the last time the server S_j was used as a cache until the associated viewing time t_j and a second test storage factor associated with storing the program at the server S_j from the last time the server S_j was on a transmission path to a previous user U_{j-1} until the associated viewing time t_j. As the skilled artisan will appreciate, for the first user U_1 and first server S_1, the storage factor will be zero.

Next, at block 92, the method determines the aggregate of the transmission factor plus the lower of the two storage time factors determined at block 90. From block 92, the method moves to decision block 94 to determine whether $S_j = S_m$. If not, the method loops back

to block 86 to consider the next server S_{j+1}, i.e., the method repeats blocks 88-94 for the server S_{j+1}.

Otherwise, the method moves to block 96 and designates the server having the lowest aggregate transmission factor and storage factor as the j^{th} server, $j \in \{1,n\}$. In other words, the j^{th} server as determined at block 96 is the server which is to cache the program and deliver the program to the user U_j under consideration. It is contemplated by the present invention that one server S_j can be designated as the server for delivering the program to more than one user U.

From block 96, the method moves to decision block 98 and determines whether $U_j = U_n$. If not, the method loops back to block 84 and repeats blocks 86-98 for the next user U_{j+1}. Otherwise, the method moves to block 100 and terminates.

While the particular system for multimedia information delivery as herein shown and described in detail is fully capable of achieving the above-stated objects of the invention, it is to be understood that it is illustrative of but one preferred embodiment, that other embodiments may exist that will be obvious to those skilled in the art, and that the scope of the present invention accordingly is to be limited by nothing other than the appended claims.

WHAT IS CLAIMED IS:

1. In a multimedia system including a plurality of program servers, a method for delivering a multimedia program having a primary transmission time to first and second users at respective user-selected first and second times, wherein the first time is simultaneous with or subsequent to the primary transmission time and the second time is subsequent to the first time, and wherein the method comprises the steps of:

(a) for each server, determining a first transmission factor associated with transmitting the program from the server to the first user;

(b) for each server, determining a first storage factor associated with storing the program at the server from the primary transmission time to the first time;

(c) designating as a first server the server having the lowest combined first transmission factor and first storage factor for transmitting the program to the first user at the first time;

(d) for each server, determining a second transmission factor associated with transmitting the program from the server to the second user;

(e) for each server, determining a second storage factor associated with storing the program at the server from the first time to the second time; and

(f) designating as a second server the server having the lowest combined second transmission factor and second storage factor for transmitting the program to the second user at the second time.

2. The method of Claim 1, further comprising the step of:

(g) storing the program in the second server at the first time.

3. The method of Claim 2, wherein the system transmits the program to a third user at a third time subsequent to the second time, and the method further comprises the steps of:

(h) for each server, determining a third transmission factor associated with transmitting the program from the server to the third user;

(i) for each server, determining a third storage factor associated with storing the program at the server from the second time to the third time; and

(j) designating as a third server the server having the lowest combined third transmission factor and third storage factor for transmitting the program to the third user at the third time.

4. The method of Claim 3, further comprising the step of:

(j) storing the program in the third server at the first or second time.

5. The method of Claim 1, wherein the system delivers the multimedia program via personal wireless communications.

6. A multimedia system for delivering a multimedia program having a primary transmission time to first and second users at respective user-selected first and second times, wherein the first time is simultaneous with or subsequent to the primary transmission time and the second time is subsequent to the first time, comprising:

a plurality of program servers;

means for determining, for each server, a first transmission factor associated with transmitting the program from the server to the first user;

means for determining, for each server, a first storage factor associated with storing the program at the server from the primary transmission time to the first time;

means for designating as a first server the server having the lowest combined first transmission factor and first storage factor for transmitting the program to the first user at the first time;

means for determining, for each server, a second transmission factor associated with transmitting the program from the server to the second user;

means for determining, for each server, a second storage factor associated with storing the program at the server from the first time to the second time; and

means for designating as a second server the server having the lowest combined second transmission factor and second storage factor for transmitting the program to the second user at the second time.

7. The system of Claim 6, further comprising:

means for storing the program in the second server at the first time.

8. The system of Claim 6, wherein the system transmits the program to a third user at a third time subsequent to the second time, and the system further comprises:

means for determining, for each server, a third transmission factor associated with transmitting the program from the server to the third user;

means for determining, for each server, a third storage factor associated with storing the program

at the server from the second time to the third time; and

means for designating as a third server the server having the lowest combined third transmission factor and third storage factor for transmitting the program to the third user at the third time.

9. The system of Claim 8, further comprising:

means for storing the program in the third server at the first or second time, wherein the system delivers the multimedia program via personal wireless communication to at least one user.

10. A method for delivering a multimedia program to a plurality of users at respective user-defined viewing times from a network having a plurality of servers, each server including a cache for storing the program with at least one of the servers being in the transmission path between another server and a user, comprising the steps of:

for the user having the second-earliest associated viewing time:

(a) for each server that can be used as a cache for the user, determining a transmission factor associated with transmitting the program from the server to the user;

(b) for each server that can be used as a cache for the user, determining the lower of a first

test storage factor associated with storing the program at the server from the last time the server was used as a cache until the associated viewing time and a second test storage factor associated with storing the program at the server from the last time the server was on a transmission path to a previous user until the associated viewing time, and selecting as a storage factor the lower of the first test storage factor and second test storage factor;

(c) designating as a user server the server having the lowest combined transmission factor and storage factor for transmitting the program to the user at the associated viewing time; and

(d) repeating steps (a)-(c) for each subsequent user, in chronological order of the associated viewing times.

11. A multimedia data delivery system for delivering a multimedia program to a plurality of users, comprising:

at least one wide area server for transmitting the multimedia program; and

a plurality of network servers for receiving the multimedia program from the wide area server and for selectively caching the multimedia program for respective storage time periods and transmitting the multimedia program at the expirations of the respective time periods to the respective users.

12. The system of Claim 11, wherein the transmission of a program from a server has an associated predetermined transmission factor, and each storage time period has an associated predetermined caching factor increasing over time and an associated preselected end time, such that the transmission factors and caching factors associated with delivering the program from the metropolitan area server to each of the users at the respective preselected times establishes an aggregate delivery factor, and the system further comprises:

a scheduler for receiving the preselected times from the users and for establishing the storage time periods in response thereto to minimize the aggregate delivery factor.

13. The system of Claim 12, wherein at least some of the network servers are arranged in tandem.

14. The system of Claim 13, wherein the scheduler further comprises means for determining which network server caches the program for each user.

15. The system of Claim 14, wherein the scheduler further comprises means for causing a network server to cache the multimedia program only at one of the preselected times.

16. The system of Claim 15, wherein the scheduler comprises means for causing a plurality of networks servers to cache the program simultaneously.

17. In a multimedia delivery network for delivering a multimedia program to a plurality of users, a network architecture characterized by:

at least one wide area multimedia program transmitter;

at least one network switch in tandem with the wide area multimedia program transmitter for relaying the multimedia program to a user; and

a network storage device associated with the network switch for caching the multimedia program for a storage time period and for causing the switch to retransmit the multimedia program to a user at a preselected time.

18. The network of Claim 17, further comprising a plurality of network switches with associated network storage devices arranged in tandem.

19. The network of Claim 18, further comprising a scheduler for receiving the preselected times from the users and for establishing the storage time periods in response thereto.

20. The network of Claim 19, wherein the transmission of the multimedia program from the multimedia program transmitter has an associated predetermined transmission factor, transmission of the multimedia program from a network switch has an associated predetermined transmission factor, and each storage time period has an associated predetermined caching factor increasing over time, such that the transmission factors and caching factors associated with delivering the program from the wide area multimedia transmitter to each of the users at the preselected times establishes an aggregate delivery factor, and the scheduler establishes the storage time periods to minimize the aggregate delivery factor.

21. The network of Claim 20, wherein the scheduler further comprises means for determining which network storage device caches the program for each user.

22. The network of Claim 21, wherein the scheduler further comprises means for causing a network storage device to cache the multimedia program only at one of the preselected times.

23. The network of Claim 22, wherein the scheduler comprises means for causing a plurality of network storage devices to cache the program simultaneously.

24. A method for delivering a multimedia program to a plurality of users at user-selected times, comprising the steps of:

(a) transmitting the program from a wide area transmitter;

(b) receiving the program at a plurality of network servers; and

(c) causing each network server to retransmit the program at one or more of the respective preselected times.

25. The method of Claim 24, further comprising the step of:

(d) causing one or more network servers to cache the program for an associated storage time period.

26. The method of Claim 25, wherein at least one of the network servers receives the program from the wide area transmitter and retransmits the program at one of the preselected times to another network server.

27. The method of Claim 26, wherein the transmission of the multimedia program from the wide area transmitter has an associated predetermined transmission factor, transmission of the multimedia program from a network server has an associated predetermined transmission factor, and each storage time period has an

associated predetermined caching factor increasing over time, such that the transmission factors and caching factors associated with delivering the program from the wide area transmitter to each of the users at the preselected times establishes an aggregate delivery factor, and the method further comprises the step of:

(e) establishing the storage time periods to minimize the aggregate delivery factor.

28. The method of Claim 27, further comprising the steps of:

(f) determining which network server caches the program for each user; and

(g) causing a network server to cache the multimedia program only at one of the preselected times.

FIG. 1

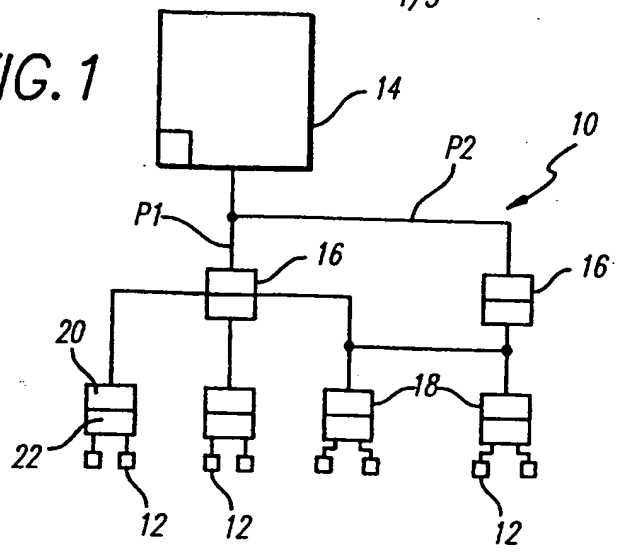


FIG. 3

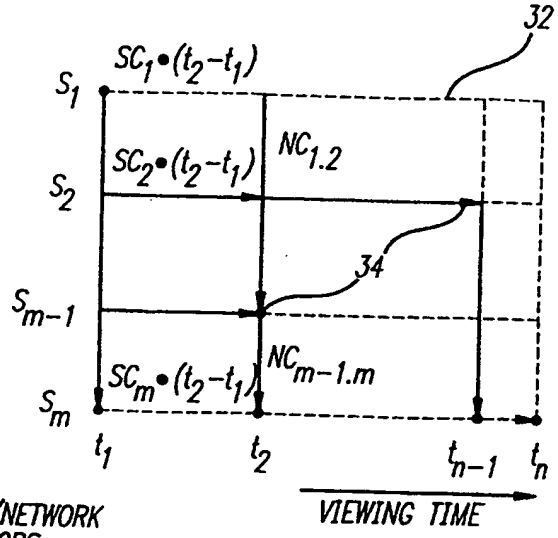


FIG. 2

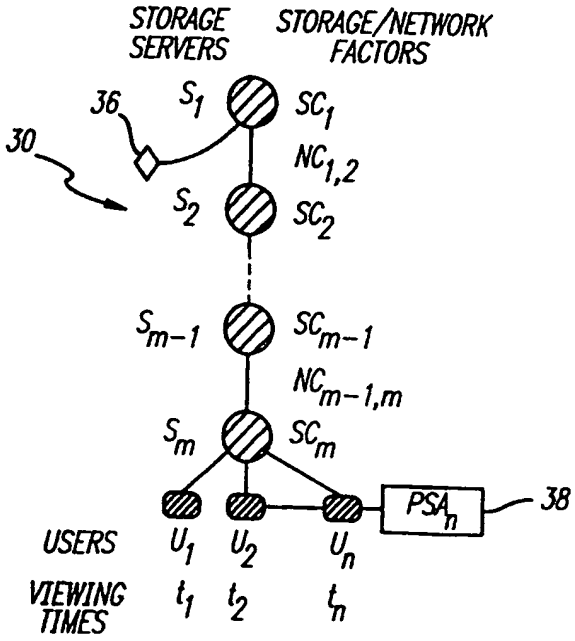


FIG. 4(A)

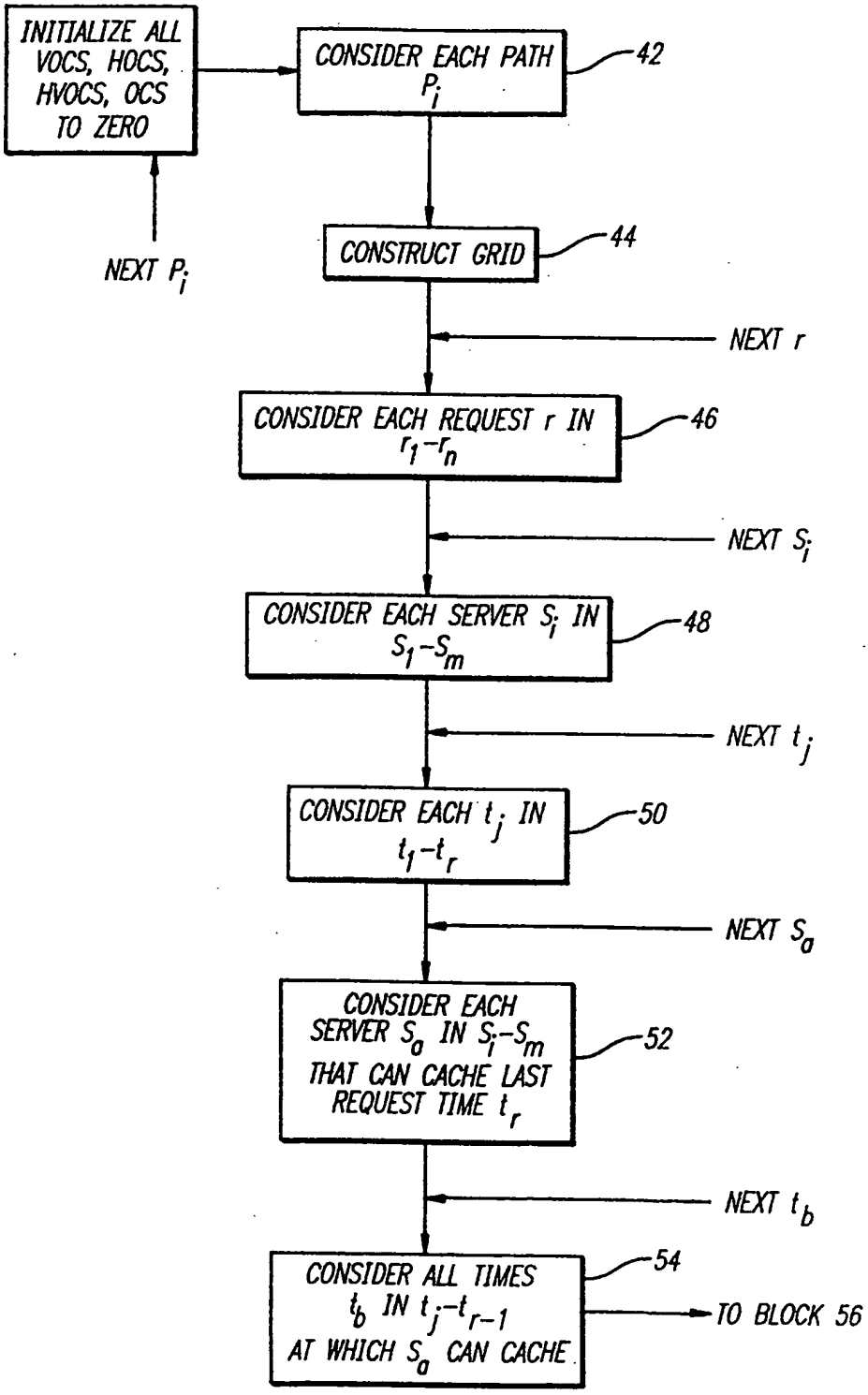

INITIALIZE ALL VOCS, HOCS, HVOCS, OCS TO ZERO
CONSIDER EACH PATH P_i
42
NEXT P_i
CONSTRUCT GRID
44
NEXT r
CONSIDER EACH REQUEST r IN r_1-r_n
46
NEXT S_i
CONSIDER EACH SERVER S_i IN S_1-S_m
48
NEXT t_j
CONSIDER EACH t_j IN t_1-t_r
50
NEXT S_a
CONSIDER EACH SERVER S_a IN S_i-S_m THAT CAN CACHE LAST REQUEST TIME t_r
52
NEXT t_b
CONSIDER ALL TIMES t_b IN t_j-t_{r-1} AT WHICH S_a CAN CACHE
54
TO BLOCK 56

FIG. 4(B)

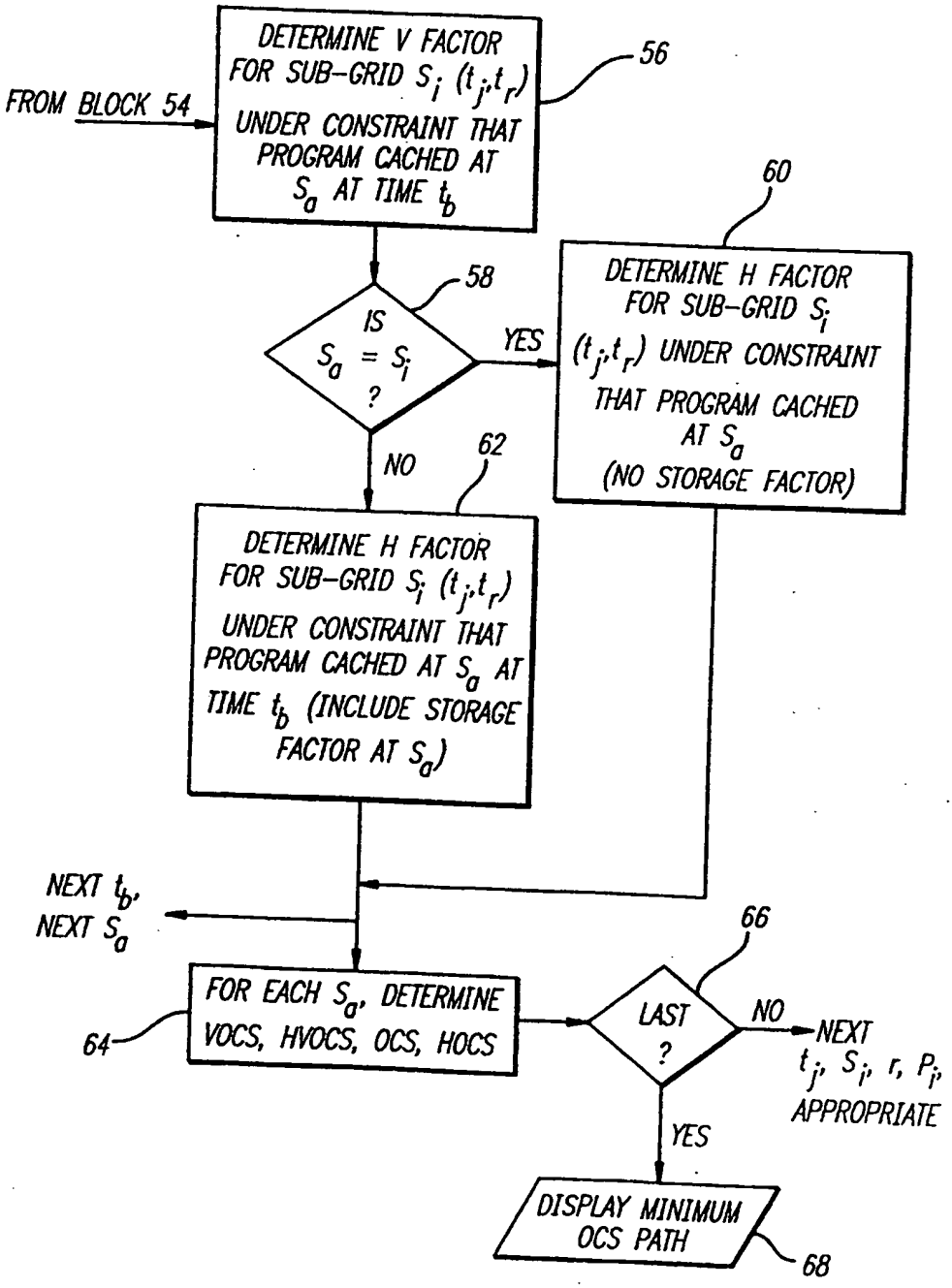

FROM BLOCK 54
DETERMINE V FACTOR FOR SUB-GRID S_i (t_j,t_r) UNDER CONSTRAINT THAT PROGRAM CACHED AT S_a AT TIME t_b
56
60
DETERMINE H FACTOR FOR SUB-GRID S_i (t_j,t_r) UNDER CONSTRAINT THAT PROGRAM CACHED AT S_a (NO STORAGE FACTOR)
58
IS $S_a = S_i$?
YES
NO
62
DETERMINE H FACTOR FOR SUB-GRID S_i (t_j,t_r) UNDER CONSTRAINT THAT PROGRAM CACHED AT S_a AT TIME t_b (INCLUDE STORAGE FACTOR AT S_a)
NEXT t_b, NEXT S_a
66
FOR EACH S_a, DETERMINE VOCS, HVOCS, OCS, HOCS
64
LAST ?
NO
NEXT t_j, S_i, r, P_i, APPROPRIATE
YES
DISPLAY MINIMUM OCS PATH
68

FIG. 5

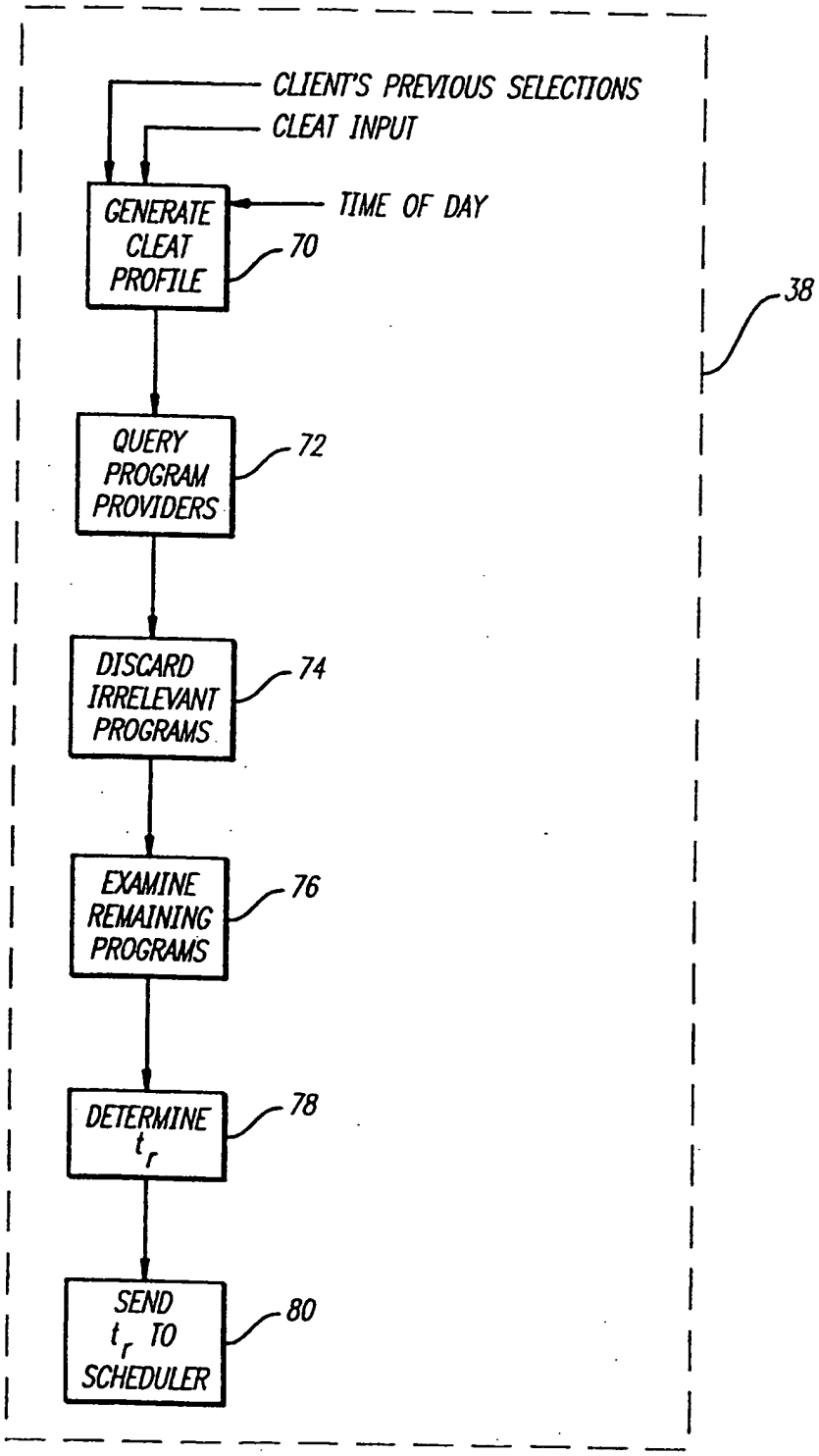
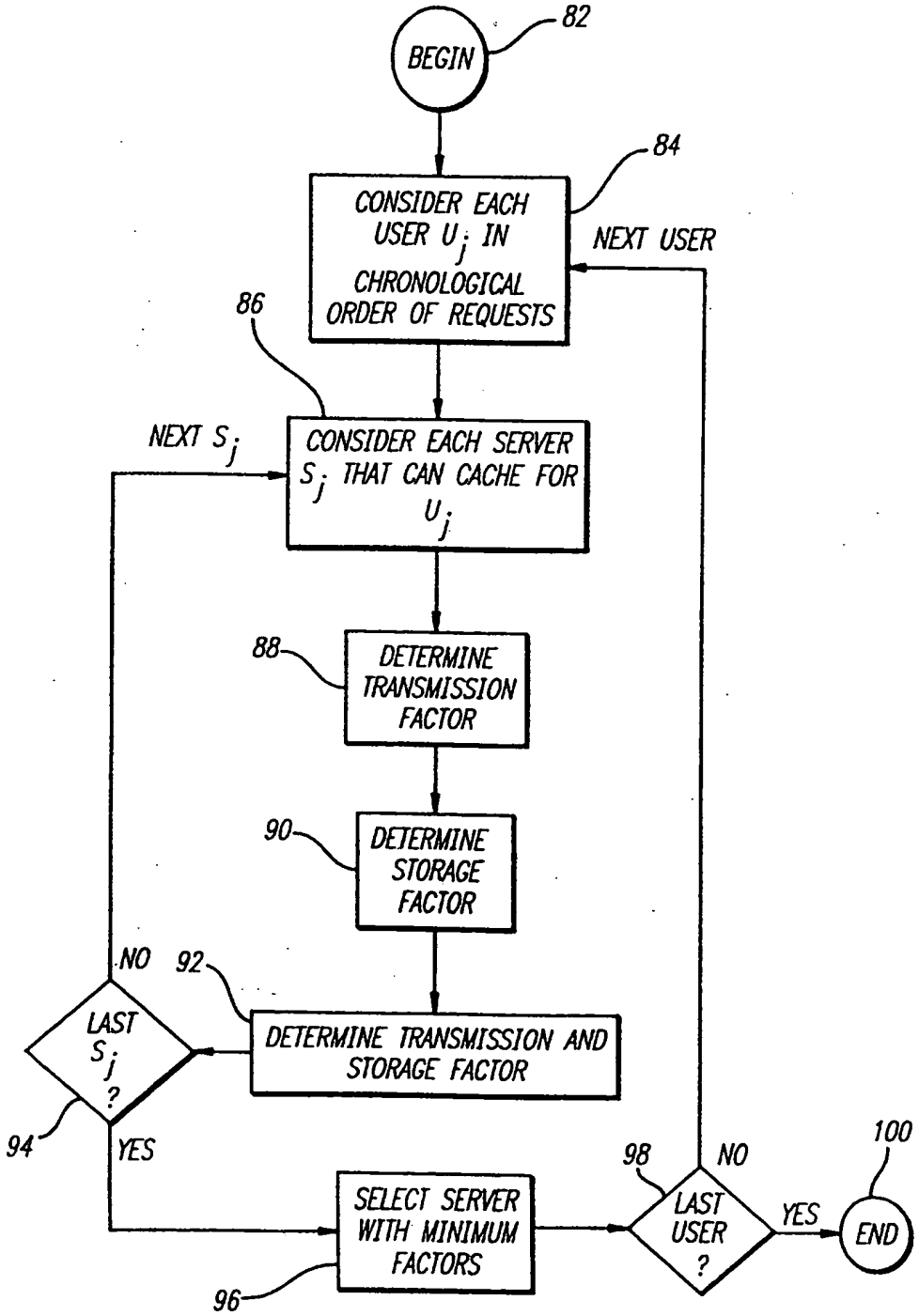

CLIENT'S PREVIOUS SELECTIONS
CLEAT INPUT
GENERATE CLEAT PROFILE
TIME OF DAY
70
38
QUERY PROGRAM PROVIDERS
72
DISCARD IRRELEVANT PROGRAMS
74
EXAMINE REMAINING PROGRAMS
76
DETERMINE t_r
78
SEND t_r TO SCHEDULER
80

FIG. 6


BEGIN
82
CONSIDER EACH USER U_j IN CHRONOLOGICAL ORDER OF REQUESTS
84
NEXT USER
86
CONSIDER EACH SERVER S_j THAT CAN CACHE FOR U_j
NEXT S_j
88
DETERMINE TRANSMISSION FACTOR
90
DETERMINE STORAGE FACTOR
92
DETERMINE TRANSMISSION AND STORAGE FACTOR
LAST S_j ?
NO
YES
94
SELECT SERVER WITH MINIMUM FACTORS
96
LAST USER ?
98
NO
YES
END
100

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01566

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G06F 15/16, 17/00, 17/30
US CL :395/200, 600
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 395/200, 600; 348/7

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS
s multimedia and server# and schedul###

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	IEEE Globecom '91 , 1991, Hardt-Kornacki et al, "Optimzation Model for the Delivery of Interactive Multimedia Documents", p.0669-0673. see abstract, p.0669 and 0670.	1-28
Y	IEEE Commnications Magazine, July 1992, Rangan et. al., "Designing an On-Demand Multimedia Service", pp.56-64, see fig.1,2,6, p.62-63	1-28
A	ACM Multimedia, June 1993, Little et. al. "A Digital On-Demand Video Service Supporting Content-Based Queries", pp.427-536.	1-28
Y	IEEE, 1991, Gelman et al. "A Store-and-Forward Architecture for Video-on-Demand Service", pp.0842-0846,	1-28

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be part of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search 29 APRIL 1995	Date of mailing of the international search report 16 JUN 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer THOMAS LEE Telephone No. (703) 305-9600

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01566

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E,Y	US, A, 5,404,505 (LEVINSON) 04 April 1995, see abstract, fig.3 and 9	1-28
Y	US, A, 5,220,420 (HOARTY et al.) 15 June 1993, see abstract.	1-28